Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning

Yue Ning,* Sathappan Muthiah,* Huzefa Rangwala,† Naren Ramakrishnan,*
* Discovery Analytics Center, Department of Computer Science, Virginia Tech, Arlington, VA 22203
† Department of Computer Science, George Mason University, Fairfax, VA 22030
* yning, sathap1, naren@vt.edu
† rangwala@cs.gmu.edu

ABSTRACT

Forecasting large-scale societal events like civil unrest movements, disease outbreaks, and elections is an important and challenging problem. From the perspective of human analysts and policy makers, forecasting algorithms must not only make accurate predictions but must also provide supporting evidence, e.g., the causal factors related to the event of interest. We develop a novel multiple instance learning based approach that jointly tackles the problem of identifying evidence-based precursors and forecasts events into the future. Specifically, given a collection of streaming news articles from multiple sources we develop a nested multiple instance learning approach to forecast significant societal events such as protests. Using data from three countries in Latin America, we demonstrate how our approach is able to consistently identify news articles considered as precursors for protests. Our empirical evaluation demonstrates the strengths of our proposed approach in filtering candidate precursors, in forecasting the occurrence of events with a lead time advantage and in accurately predicting the characteristics of civil unrest events.

Keywords

Multi-Instance Learning, Event Detection, Text Mining

1. INTRODUCTION

Forecasting societal uprisings such as civil unrest movements is an important and challenging problem. Open source data (e.g., social media and news feeds) have been proven to serve as surrogates in forecasting a broad class of events, e.g., disease outbreaks [1], election outcomes [18, 22], stock market movements [6] and protests [19]. While many of these works focus on predictive performance, there is a critical need to develop methods that also yield insight by identifying precursors to events of interest.

This paper focuses on the problem of identifying precursors (evidence) for forecasting significant societal events, specifically protests. Modeling and identifying the precursors for a given protest is useful for human analysts and policy makers as it discerns the underlying reasons behind the civil unrest movement. In particular, the objective of this paper is to study and forecast protests across different cities in three Latin American countries (Argentina, Brazil and Mexico). 6000 news outlets are tracked daily across these countries with the goal of forecasting protest occurrences with at least one day of lead time. From the news feeds, we also aim to identify the specific news articles that can be considered as precursors for the target event.

Figure 1 shows an example of precursors identified by our model. On the right of the timeline is a news report about a protest event in Argentina. The connected dots denotes the generated probabilities of a protest event over the days leading upto this protest. From this example, we find that within 10 days before the event, there are multiple precursor events identified as highly probable leading indicators of a protest. Most significant societal events are a consequence of several factors that affect different entities within communities and their relationships with each other (or the government) over time. In this specific example, the leading precursor was an article commenting on standards of living in Argentina and rising poverty levels. The International Court of Justice also delivered a verdict on the debt crisis. All these factors led to the final protest involving the general population across the country demanding better work opportunities.

We formulate the precursor identification and forecasting problem in a novel multiple instance learning algorithm (MIL) setting. Multiple instance learning algorithms [3, 28] are a class of supervised learning techniques that accept labels for groups of instances, but where labels for individual instances are not avaialble. In our formulation, instances denote news articles and while class labels are not associated with individual news articles, a group of news articles is attached with a label (indicating the occurrence of a protest). We further extend the standard MIL formulation by introducing a nested structure, wherein we group news articles published in a given day at the first level and then group the collection of individual days at the second level. This nested MIL approach allows for modeling the sequential constraints between the news articles (grouped by days) published on different days and also provides a probabilistic estimate for every news article and the collection of news article. This estimate is significant because it indicates for a given news article the probability of it signaling a protest event. Recall that in our datasets we do not have any training labels to associate a protest per news article.
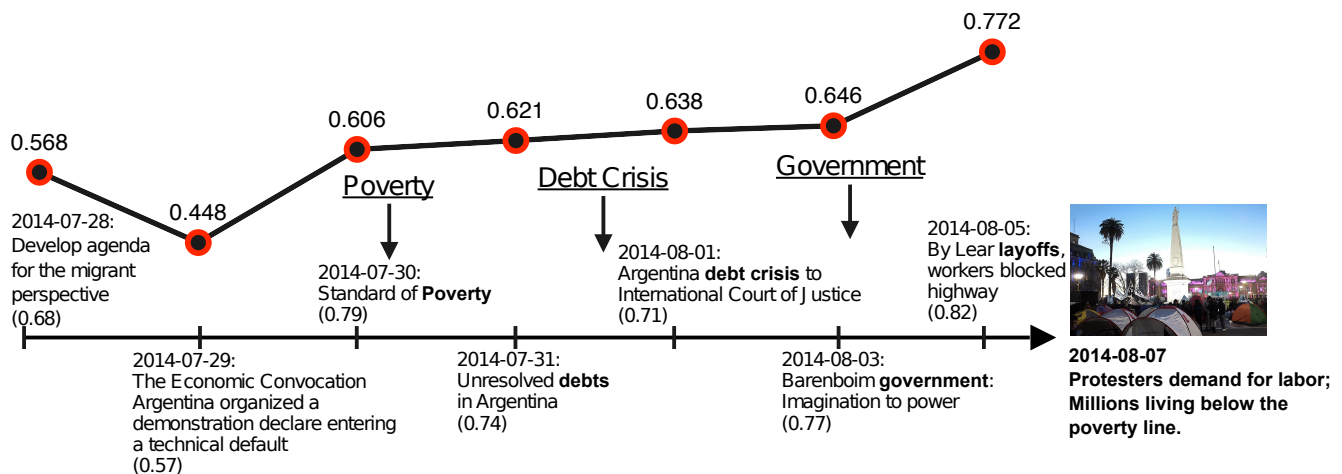
Figure 1: Precursor story line for a protest event in Argentina. The x-axis is the timeline. The dots above with numbers are the probabilities for each day that the model generated for the target event. Each precursor document is titled in the timeline.

The main contributions of this study are summarized as follows:

1. **A novel nested framework of multi-instance learning for event forecasting and precursor mining.** We formulate event forecasting and precursor mining for multiple cities in a country as a multi-instance learning problem with a nested structure. By estimating a prediction score for each instance in the history data, we automatically detect significant precursors for different events.

2. **Harness temporal constraints in multi-instance learning.** We explore different penalty function and regularizations where we employ the temporal information in our dataset under assumption that most events of interest are follow-up reports of other events that happened before, and most planned events are developing over time.

3. **Modeling for various event categories in multiple geo-locations**. We extend the nested MIL formulation for general purpose multi-class classification to determine necessary attributes of events in terms of their underlying population.

4. **Application and evaluation with comprehensive experiments.** We evaluate the proposed methods using news data collected from July 2012 to December 2014 in three countries of Latin America: Argentina, Mexico, and Brazil. For comparison, we implement other multi-instance algorithms, and validate the effectiveness and efficiency of the proposed approach. We also perform qualitative and quantitative analysis on the precursors inferred by our model.

The rest of this paper is organized as follows. We discuss related work in Section 2. Section 3 introduces the problem setup and our proposed model based on multi-instance learning is presented in Section 4. This section is then followed by experiments and evaluations on real world datasets, presented in Section 5 and Section 6 . Finally, we conclude with a summary of the research in Section 7.

2. RELATED WORK

Event Detection and Forecasting. Event detection and forecasting from online open source datasets has been an active area of research in the past decade. Both supervised and unsupervised machine learning techniques have been developed to tackle different challenges. Linear regression models use simple features to predict the occurrence time of future events [4, 6, 10, 18]. Advanced techniques use a combination of sophisticated features such as topic related keywords, as input to support vector machines, LASSO and multi-task learning approaches [23, 20]. Ramakrishnan *et al.* [19] designed a framework (EMBERS) for predicting civil unrest events in different locations by using a wide combination of models with heterogeneous input sources ranging from social media to satellite images. Zhao *et al.* [27] combine multi-task learning and dynamic features from social networks for spatial-temporal event forecasting. Generative models have also been used in [26] to jointly model the temporal evolution in semantics and geographical burstiness within social media content. Laxman *et al.* [13] designed a generative model for categorical event prediction in event streams using frequent episodes. However, few existing approaches provide evidence and interpretive analysis as support for event forecasting.

Identifying Precursors. Identifying precursors for significant events is an interesting topic and has been used extensively for interpretive narrative generation and in storytelling algorithms [11]. Rong *et al.* [21] developed a combinational mixed Poisson process (CMPP) model to learn social, external and intrinsic influence in social networks.

Multiple Instance Learning. In the multiple instance learning (MIL) paradigm, we are given labels for sets of instances commonly referred as *bags* or *groups*. However, individual instance-level labels are unknown or missing. The bag-level labels are assumed to be an association function (e.g., OR, average) of the unknown instance level labels. One approach to MIL adapts support vector machines (SVMs) by: (i) modifying the maximum margin formulation to discriminate between bags rather than individual instances [3], and (ii) developing kernel functions that operate directly on bags [9]. Other multiple instance learning approaches

Table 1: Notation.

Variable	Meaning
$\mathcal{S} = \{\mathbb{S}\}$	a set of n "super bags" in our dataset
$\mathbb{S} = [\mathcal{X}_i], i \in \{1, ..., t\}$	an ordered set of t "bags" in $\mathbb{S}$
$\mathcal{X}_i = \{\mathbf{x}_{ij}\}, j \in \{1, ..., n_i\}$	a set of instances with $n_i = |\mathcal{X}_i|$, number of instance in a bag $\mathcal{X}_i$
$\mathbf{x}_{ij} \in \Re^{V \times 1}$	the j-th instance in set $\mathcal{X}_i$, a V-dimension vector
$Y \in \{-1, +1\}$	label of super bag
$P \in [0, 1]$	estimated probability for a super bag
$\mathbb{P}_i \in [0, 1]$	the probability of bag i in super bag to be positive
$p_{ij} \in [0, 1]$	the probability of an instance $\mathbf{x}_{ij}$ in bag $\mathcal{X}_i$ in super bag to be positive
$C \in \{1, 2, ..., K\}$	multi-class label of super bag

and various applications are found in a detailed survey [2]. Specifically, the generalized MIL [25] formulation assumes the presence of multiple concepts and a bag is classified as positive if there exists instances from every concept. Relevant to our work, besides predicting bag labels, Liu *et al.* [15] seek to identify the key instances within the positively-labeled bags using nearest neighbor techniques. Recent work [12] has focused on instance-level predictions from group labels (GICF) and allowed for the application of general aggregation functions with applications to detecting sentiments associated with sentences within reviews.

The methods proposed in this paper can be viewed as complementary to prior work, casting the forecasting and precursor discovery problems within novel extensions of multiple instance learning.

3. PROBLEM FORMULATION

Given a collection of streaming media sources (e.g., news feeds, blogs and social network streams), the objective of our study is to develop a machine learning approach to forecast the occurrence of an event of interest in the near future. Specifically, we focus on forecasting protests or civil unrest movements in Latin America from a daily collection of published news articles. Besides forecasting the protest, we aim to identify the specific news articles from the streaming news outlets that can be considered as supporting evidence for further introspection by an intelligence analyst. We refer to these identified articles as *precursors* for a specific protest.

Figure 2 provides an overview of our proposed approach and problem formulation. Here, we show groups of news articles collected daily, five days prior to the specific protest event (being forecast). Within our proposed MIL-based formulation, each news article is an individual instance, the collection of news articles published on a given day is a bag, and the ordered collection of bags (days) is denoted as a super-bag (explained in detail later). For this study, each individual news article is represented by a distributed representation for text derived using a framework such as text embedding [14]. Figure 2 shows that for certain days within the collection we attempt to identify news articles (highlighted) that are considered as precursors from the entire collection of input news articles used for forecasting the occurrence of a specific target.
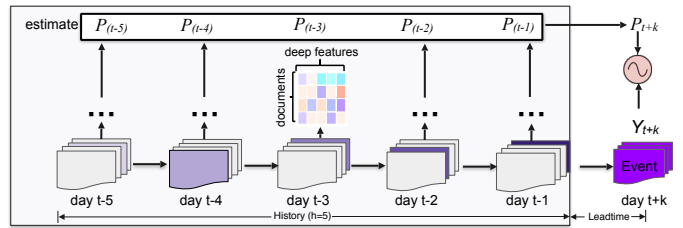

Figure 2: Overview of proposed approach to forecasting and precursor discovery.

3.1 Formal Definition and Notation

For a given protest event e occurring on day $t + k$, we assume that for each day before the event we are tracking a multitude of news sources. We represent the collection of n_i news articles published on a given day i by $\mathcal{X}_i = \{\mathbf{x_{i,1}} \ldots \mathbf{x_{i,n_i}}\}$, where the j-th news article is represented by $\mathbf{x_{ij}}$. The ordered collection of news articles for the protest event up to day t can be represented as a super-bag, $\mathbb{S}_{1:t} = \{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$. The occurrence of the protest event at time $t + k$ is denoted by $\mathcal{Y}_{t+k} \in \{-1, +1\}$ where 1 denotes a protest and -1, otherwise.

The forecasting problem can be formulated as learning a mathematical function $f(\mathbb{S}_{1:t}) \to \mathcal{Y}_{t+k}$ that maps the input, an ordered collection of news articles extracted per day to a protest indicator k days in the future from the day t. To identify the news articles considered as precursors (evidence), we aim to estimate a probability for each news article on any given day that signifies the occurrence of a given protest. For a news article $\mathbf{x_{ij}}$, we denote this estimated probability value by p_{ij}. As such, given the collection of news articles we identify the precursor set as the ones with p_{ij} greater than a fixed threshold τ. We represent this precursor set of documents as a subset of the original super-bag, given by $\{\mathbf{x_{ij}} \in \mathbb{S}_{1:t} \mid p_{ij} > \tau\}$. As a secondary objective, we aim to forecast the occurrence of an event with a long lead time i.e., large values of k. Table 1 captures the notation and definitions used in this study.

4. METHODS

We first provide our intuition behind formulating the precursor discovery and forecasting problem within a novel extension of multiple instance learning algorithm. Parallel to the standard multiple instance learning algorithms we have a group of news articles (bags) with labels available only for the entire bag (i.e., leading to a protest); and one of the objectives is to train a classifier to predict the bag-level label. In addition to predicting the group-level labels, we also care about predicting the labels for individual news articles (instances) since they signify the precursor. Various MIL formulations extend the basic definition with a similar motivation, i.e., to estimate the key instances within a bag or provide instance-level labels. However, our problem setting has a two-level grouping structure with sequential constraints, i.e., we capture news articles per day (bags) and group the days to form a super-bag with labels only available at the super-bag level. As such, we propose a nested multiple instance learning formulation for predicting the super bag level labels (forecast) and then estimate the bag-level and instance-level probabilities for identifying association of the bag and instance with the event, respectively. We de-

veloped various extensions of our proposed approach to tie the different sequential and group constraints.

4.1 Nested MIL model (nMIL)

We model the instance level probability estimates p_{ij} for a news article j on day i to associate with a targeted event e with a logistic function. These probability estimates indicate how related the specific instance is to the target event, e. Higher the probability value, the more related the document is to the target event and most probably represents a precursor that contains information about causes of the target event.

$$p_{ij} = \sigma(\mathbf{w}^T \mathbf{x}_{ij}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_{ij}}}. \tag{1}$$

Here, $\mathbf{w}$ denotes the learned weight vector for our model. The probability for a day (or bag) is then modeled as the average of probability estimates of all instances in a day [12]. Hence, for each bag:

$$\mathbb{P}_i = \mathcal{A}(\mathcal{X}_i, \mathbf{w}) = \frac{1}{n_i} \sum_{j}^{n_i} p_{ij}, \tag{2}$$

where $\mathcal{A}$ is an aggregation function.

We then model the probability of a super-bag $\mathbb{S}$ (associated with an event e) being positive as the average of the probability of all t bags within the super bag to be positive (related to the target event). Thus:

$$P = \mathcal{A}(\mathbb{S}, \mathbf{w}) = \frac{1}{t} \sum_{i}^{t} \mathbb{P}_i \tag{3}$$

For a given super bag $\mathbb{S}$, as all the t bags within it are temporally ordered, the probability estimates for a given bag (day) is assumed to be similar to its immediate predecessor. This consistency in consecutive bag probabilities is modeled by minimizing the following cross-bag cost as below:

$$g(\mathcal{X}_i, \mathcal{X}_{i-1}) = (P_i - P_{i-1})^2 \tag{4}$$

Finally, given a set of true labels Y for the super bags, we can train our model by minimizing the following cost function w.r.t to $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{\beta}{n} \sum_{\mathbb{S} \in \mathcal{S}} f(\mathbb{S}, Y, \mathbf{w}) + \frac{1}{n} \sum_{\substack{\mathbb{S} \in \mathcal{S}; \\ \mathcal{X}_i, \mathcal{X}_{i-1} \in \mathbb{S}}} \frac{1}{t} \sum_{i=1}^{t} g(\mathcal{X}_i, \mathcal{X}_{i-1}, \mathbf{w}) \tag{5}$$

$$+ \frac{1}{n} \sum_{\substack{\mathbb{S} \in \mathcal{S}; \mathcal{X}_i \in \mathbb{S} \\ \mathbf{x}_{ij} \in \mathcal{X}_i}} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n_i} \sum_{j=1}^{n_i} h(\mathbf{x}_{ij}, \mathbf{w}) + \lambda R(\mathbf{w})$$

Here,
- $f(\mathbb{S}, Y, \mathbf{w}) = -\mathbf{I}(Y = 1)logP - \mathbf{I}(Y = -1)(log(1 - P))$ is the negative log-likelihood function that penalizes the difference between prediction and the true label for super bag $\mathbb{S}$ where $\mathbf{I}(\cdot)$ is the indicator function.
- $g(\mathcal{X}_i, \mathcal{X}_{i-1}, \mathbf{w})$ is the cross-bag cost defined in Equation. 4
- $h(\mathbf{x}_{ij}, \mathbf{w}) = max(0, m_0 - sgn(p_{ij} - p_0)\mathbf{w}^T \mathbf{x}_{ij})$ represents the instance level cost. Here, sgn is the sign function; m_0 is a crucial margin parameter used to separate the positive

and negative instances from the hyper line in the feature space; p_0 is a threshold parameter to determine positiveness of instance.
- $R(\mathbf{w})$ is the regularization function.
- β, λ are constants that control the trade-offs between the loss function and regularization function.

4.1.1 Cross-bag Similarity (nMIL$^\Delta$)

The cross-bag similarity $g(,)$ in the above equation does not allow for sudden changes in the day-level probabilities caused due to newer events happening on the current day. We update the cost function across days (bags) (Equation 4) as follows:

$$g(\mathcal{X}_i, \mathcal{X}_{i-1}) = \Delta(\mathcal{X}_i, \mathcal{X}_{i-1})(P_i - P_{i-1})^2 \tag{6}$$

The objective function above allows for label information to spread over the manifold in the feature-space. As such, we compute $\Delta(,)$ as the pairwise cosine similarity between the news articles in $\mathcal{X}_i$ and $\mathcal{X}_{i-1}$. Since we do not have ground truth labels for the bag level (day) we make this consistency assumption that estimated probabilities for consecutive days should be similar if the news articles have similarity in the feature space as well. This model is referred by **nMIL$^\Delta$** and allows for sudden changes in how events unfold.

4.2 Sequential Model (nMIL$^\Omega$)

The basic **nMIL** models assume that there exists a single weight vector across all the days (bags) within a super bag. To model the sequential characteristics of the articles published across consecutive days, we extend this formulation by learning individual weight vectors for each of the historical days. Assuming t days within a super bag $\mathbb{S}$ we learn a weight vector for each individual day represented as $\Omega = [\mathbf{w}_1, \ldots, \mathbf{w}_t]$; where $\mathbf{w}_i$ is the weight vector learned for day i. In this setting, the individual weight vectors are still learned together in a joint fashion akin to multi-task learning approaches [7]. However, the probability of a news article j on day i will be given by $p_{ij} = \sigma(\mathbf{w}_i^T \mathbf{x}_{ij})$. This formulation is called **nMIL$^\Omega$** and given by:

$$J(\mathbf{\Omega}) = \underbrace{\frac{\beta}{n} \sum_{\mathbb{S} \in \mathcal{S}} f(\mathbb{S}, \mathbf{\Omega}, Y)}_{\text{empirical loss}} + \underbrace{\frac{1}{n} \sum_{\substack{\mathbb{S} \in \mathcal{S}; \\ \mathcal{X}_i, \mathcal{X}_{i-1} \in \mathbb{S}}} \frac{1}{t} \sum_{i=1}^{t} g(\mathcal{X}_i, \mathcal{X}_{i-1}, \mathbf{w}_i)}_{\text{sequential loss}}$$

$$+ \underbrace{\frac{1}{n} \sum_{\substack{\mathbb{S} \in \mathcal{S}; \mathcal{X}_i \in \mathbb{S} \\ \mathbf{x}_{ij} \in \mathcal{X}_i}} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n_i} \sum_{j=1}^{n_i} h(\mathbf{x}_{ij}, \mathbf{w}_i) + \lambda R(\mathbf{\Omega})}_{\text{unsupervised loss}} \tag{7}$$

Just like the multi-task learning algorithms, the regularization term $R(\mathbf{\Omega})$ can be modified to capture the various relationship-based constraints. However, in this study we ignore these specialized approaches focusing only on the MIL paradigm.

4.3 Multiclass Classification

We also extend our developed **nMIL** formulations to solve general purpose multiclass classification problems rather than binary classification problems. Within our domain, each labeled event is manually attached with event population.

Event population indicates the size/community of people who participated in the protest event.

For multiclass classification problems, we train one-versus-rest classifiers for each of the classes learning a separate weight vector per class. When classifying a super bag to a specific event population we first forecast the binary protest indicator label for a super bag. Next, we apply the multiclass classification only on the predicted positive examples.

4.4 Optimization

We perform online stochastic gradient decent optimization to solve our cost function and test our model on new data to predict super bag label. For every iteration in our algorithm, we randomly choose a super-bag $(\mathbb{S}, Y)$ from the training dataset $\mathcal{S}$ by picking an index $r \in \{1, \ldots, n\}$ using a standard uniform distribution. Then we optimize an approximation based on the sampled super-bag by:

$$J(\mathbf{w}; \mathbb{S}) = \beta f + \frac{1}{t}\sum_i^t g_i + \frac{1}{t}\sum_i^t \frac{1}{n_i}\sum_j^{n_i} h_{ij} + \lambda R(\mathbf{w}) \quad (8)$$

The gradient of the approximate function is given by:

$$\nabla J(\mathbf{w}) = \frac{\partial J(\mathbf{w}; \mathbb{S})}{\partial \mathbf{w}} = \lambda \mathbf{w} \quad (9)$$

$$- \frac{Y - P}{P(1 - P)}\frac{\beta}{t}\sum_i^t \frac{1}{n_i}\sum_k^{n_i} p_{ij}(1 - p_{ij})\mathbf{x}_{ij}$$

$$+ \frac{1}{t}\sum_i^t 2(P_i - P_{i-1})\frac{1}{n_i}\sum_j^{n_i} p_{ij}(1 - p_{ij})\mathbf{x}_{ij}$$

$$- \frac{1}{t}\sum_i^t 2(P_i - P_{i-1})\frac{1}{n_{i-1}}\sum_j^{n_{i-1}} p_{vj}(1 - p_{vj})\mathbf{x}_{vj}$$

$$- \frac{1}{t}\sum_i^t \frac{1}{n_i}\sum_j^{n_i} sgn(p_{ij} - p_0)\mathbf{x}_{ij}(o_{ij})$$

where $v = i - 1, o_{ij} = I(sgn(p_{ij} - p_0)\mathbf{w}\mathbf{x}_{ij} < m_0)$. We update the weight vector using a varied learning rate and $\mathbf{w}' = \mathbf{w} - \eta\nabla(\mathbf{w})$ using mini-batch stochastic gradient descent where η is the learning rate at current iteration.

4.5 Precursor discovery using nMIL

In the **nMIL** model, each super-bag consists of an ordered set of bags and each bag represents the documents in one day in the city for which we are forecasting a protest event. We present in Algorithm 1 the steps to identify news articles as precursors based on their estimated probability given by $p_{ij} > \tau$.

5. EXPERIMENTS

5.1 Datasets

The experimental evaluation was performed on news documents collected from around 6000 news agencies between July 2012 to December 2014 across three countries in South America, viz. Argentina, Brazil, and Mexico. For Argentina and Mexico, the input news articles were primarily in Spanish and for Brazil, the news articles were in Portuguese.

The ground truth information about protest events, called the gold standard report (GSR) is exclusively provided by MITRE [19]. The GSR is a manually created list of civil

---

Algorithm 1 Precursor Discovery in **nMIL**

1: **procedure** PD-**nMIL**
2:     **Input**: $\mathcal{S} = \{(\mathbb{S}_r, Y_r)\}_{r \in n^+}, \mathcal{M}$
3:     **Output**: $\{(ps_r, Y_r)\}_{r \in n^+}$
4:     **for** super bag (S_r, Y_r) **do**
5:         $ps_r = []$
6:         **for** t = 1,2,...,h(history days) **do**
7:             $y_t = []$
8:             **for** $\mathbf{x}_{tm} \in \mathcal{X}_t$ **do**
9:                 $\hat{y}_{tm} = \sigma(\hat{\mathbf{w}}\mathbf{x}_{tm})$
10:                 **if** $\hat{y}_{tm} > \tau$ **then**
11:                     $y_t \leftarrow (m, \hat{y}_{tm})$
12:             sort(y_t) by $\hat{y}_{tm}$ in descending order
13:             $ps_r \leftarrow$ m where m in top(y_t)
      return $\{(ps_r, Y_r)\}_{r \in n^+}$

---

unrest events that happened during the period 2012-2014. A labeled GSR event provides information about the geographical location at the city level, date, type and population of a civil unrest news report extracted from the most influential newspaper outlets within the country of interest. These GSR reports are the target events that are used for validation of our forecasting algorithm. We have no ground truth available for verifying the validity of the precursors.

Argentina: We collected data for Argentina from newspaper outlets including *Clarin* and *Lanacion* for the period of July 2010 to December 2014. There are multiple protest events in Argentina during this period. For instance, people protested against the government and utility/electricity-providing companies because of heatwaves in Dec. 2013.

Brazil: For Brazil, we obtained data from news agencies including the three leading news agencies in Brazil; *O Globo, Estadao,* and *Jornal do Brasil* from November 2012 to September 2013. During this period Brazil faced several mass public demonstrations across several Brazilian cities stemming from a variety of issues ranging from transportation costs, government corruption, and police brutality. These mass protests were initiated due to a local entity advocating for free public transportation. This period had an unusually high social media activity and news coverage and is also known as the "Brazilian Spring" [1].

Mexico: For Mexico, we tracked news agencies including the top outlets: *Jornada, Reforma, Milenio* from January 2013 to December 2014. Over 619 days, we noticed 71 news articles per day on average. There were more than 2000 protest events in this two-year period with major unrest movements in 2013 led by teachers and students demanding education reforms by protesting against the government.

5.2 Experimental Protocol

The GSR signifies the occurrence of a protest event on a given day at a specific location. To evaluate the MIL-based forecasting and precursor discovery algorithms, for each protest event we extract all the published news articles for up to 10 days before the occurrence of the specific event. This ordered collection of per-day news documents up to the protest day are considered as positive super bags. For negative samples, we identify consecutive sets of five days within our studied time periods for the different countries when no

---

[1]http://abcnews.go.com/ABC_Univision/
brazilian-spring-explainer/story?id=19472387

protest was reported by the GSR. The ordered collection of per-day news documents not leading to a protest are considered as negative super bags for the nMIL approach. For any news article (i.e., an individual instance) within a positive/negative super-bag we have no label (or ground truth). As part of the precursor discovery algorithm, we estimate a probability for an individual instance to signal a protest (by showing evidence). It is important to note that the GSR linked news article for a protest is never used for training purposes. Having identified the positive and negative samples, we split our datasets into training and testing partitions and perform 3-fold cross-validation. A single run of the model on a machine with 4 cores and 16 GB memory takes about 250 seconds.

We study the performance of forecasting models with varying lead time and varying historical days. Lead time (l) indicates the number of days in advance the model makes predictions and historical days (h) denotes the number of days over which the news articles are extracted as input to the prediction algorithms. As an example, if l is set to 1, then the model forecasts if a protest event is planned for the next day. Setting the historical days, h, to 5 denotes that we use news from five days before the current day to make the forecast. We varied l from 1 to 5 and h from 1 to 10 and trained 50 different models for the different approaches to study the characteristics of the developed approaches with varying lead time and historical days.

For event forecasting, we evaluate the performance by standard metrics including precision, recall, accuracy and F1-measure.

5.3 Comparative Approaches

We compare the proposed **nMIL** models to the following approaches:

- **SVM** : We use the standard support vector machine formulation [8] by collapsing the nested grouping structure and assigning the same label for each news article as its super-bag (for training). During the prediction phase, the SVM yields the final super-bag prediction (forecast) by averaging the predicted label obtained for each of the instances.

- **MI-SVM** [3]: The MI-SVM model extends the notion of a margin from individual patterns to bags. Notice that for a positive bag the margin is defined by the margin of the "most positive" instance, while the margin of a negative bag is defined by the "least negative" instance. In our case, we collapse the news articles from the different historical days into one bag and apply this standard MIL formulation.

- **Relaxed-MIL** (**rMIL^{nor}**) [24]: Similar to the MI-SVM baseline, we collapse the news articles into one bag. However, unlike the MI-SVM formulation the **rMIL^{nor}** model can provide a probabilistic estimate for a given document within a bag to be positive or negative.

- **Modified Relaxed-MIL** (**rMIL^{avg}**): This approach is similar to the **rMIL^{nor}** , except we compute the probability of a bag being positive by taking average of estimate of each instance in the bag rather than using the Noisy-OR model discussed above.

- **GICF** [12]: This model optimizes a cost function which parameterized the whole-part relationship between groups and instances and pushes similar items across different groups to have similar labels.

5.4 Feature Description

In practice, finding good feature representations to model the news articles is not a trivial problem. Traditionally the bag-of-words representation allows for easy interpretation but also requires pre-processing and feature selection. Several researchers have developed efficient and effective neural network representations for language models [5, 16, 17]. Specifically, we learn deep features for documents by taking advantage of the existing doc2vec model. For each document, we generate a 300 dimension vector for training with a contextual window size of 10 in an unsupervised version [14]. We compared the performance of deep features with traditional TF-IDF features but the results showed little difference. Thus, we only report the evaluation of models with deep features.

6. RESULTS AND DISCUSSION

In this section, we evaluate the performance of the proposed models. First, we evaluate the effectiveness and efficiency of the methods on real data in comparison with baseline methods on multiple configurations of forecasting tasks. Then, we study and analyze the quality of precursors with respect to quantitative and qualitative measures. Multi-class forecasting evaluation is also provided for one of the countries. Finally, we perform a sensitivity analysis of performance regarding parameters in the proposed model.

6.1 How well does nMIL forecast protests?

6.1.1 Comparative Evaluation

Table 2 reports the prediction performance of the **nMIL** approach in comparison to other baseline approaches for the task of forecasting protests. Specifically, we use set $\beta = 3.0$, $\lambda = 0.05$, $m_0 = 0.5$ and $p_0 = 0.5$ (β, λ chosen by sensitivity analysis, m_0, p_0 by default setting in hinge loss) and report the average accuracy and F1 score along with standard deviation for predicting protests across multiple runs of varying historical days with lead time set to 1. We observe that the **nMIL** approaches outperform the baseline approaches across all the three countries. The **rMIL^{nor}** approach performs poorly because the the noisy-OR aggregation function associating the bag-level labels to instance-level labels forces most of the news articles within the positive bags to have probability values close to 1. However, given the large collection of news articles available per day only a subset of them will provide a signal/evidence for a protest. For Argentina, the **nMIL** and **nMIL$^{\Delta}$**approaches outperformed the best baseline (**MI-SVM**), by **7%** and **8%** with respect the average F1 score, respectively.

Figure 3 shows the changes to F1 score for the proposed **nMIL** approach in comparison to **SVM** , **MI-SVM** and **rMIL^{avg}** for different number of historical days that are used in training with lead time set to 2. We trained 10 different models that use different number of historical days respectively varying from 1 to 10. These results show that the methods that utilize the nested structure (**nMIL** ,**nMIL$^{\Delta}$**) within the multi-instance learning paradigm generally performed better than others. Moreover, the proposed **nMIL**

Table 2: Event forecasting performance comparison based Accuracy (Acc) and F-1 score w.r.t to state-of-the-art methods. The proposed **nMIL** , **nMIL$^\triangle$**, **nMIL$^\Omega$** method outperform state-of-the-art methods across the three countries.

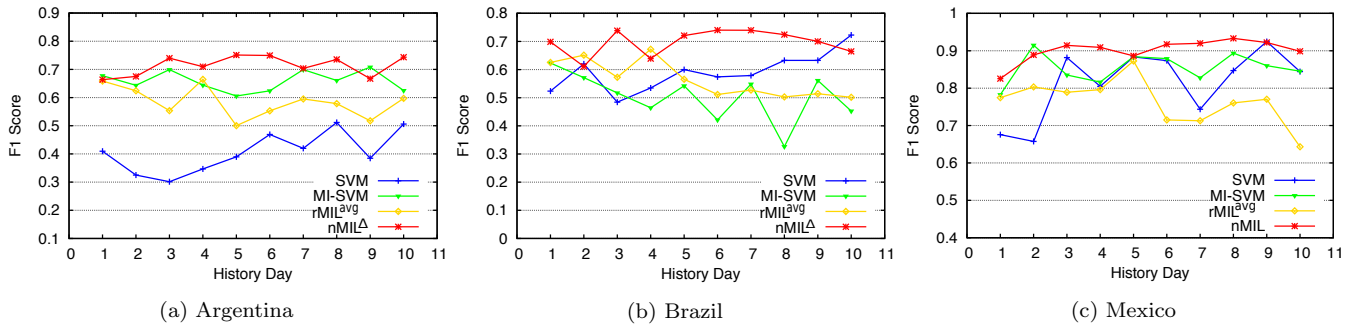|	Argentina		Brazil		Mexico	
	Acc	F-1	Acc	F-1	Acc	F-1
SVM	0.611($\pm$0.034)	0.406($\pm$0.072)	0.693($\pm$0.040)	0.598($\pm$0.067)	0.844($\pm$0.062)	0.814($\pm$0.091)
MI-SVM	0.676($\pm$0.026)	0.659($\pm$0.036)	0.693($\pm$0.040)	0.503($\pm$0.087)	0.880($\pm$0.025)	0.853($\pm$0.040)
rMIL^{nor}	0.330($\pm$0.040)	0.411($\pm$0.092)	0.505($\pm$0.012)	0.661($\pm$0.018)	0.499($\pm$0.009)	0.655($\pm$0.025)
rMIL^{avg}	0.644($\pm$0.032)	0.584 ($\pm$0.055)	0.509($\pm$0.011)	0.513($\pm$0.064)	0.785($\pm$0.038)	0.768($\pm$0.064)
GICF	0.589($\pm$0.058)	0.624($\pm$0.048)	0.650($\pm$0.055)	0.649 ($\pm$0.031)	0.770($\pm$0.041)	0.703($\pm$0.056)
nMIL	**0.709**($\pm$0.036)	0.702($\pm$0.047)	**0.723**($\pm$0.039)	0.686($\pm$0.055)	**0.898**($\pm$0.031)	**0.902**($\pm$0.030)
nMIL$^\triangle$	0.708($\pm$0.039)	**0.714**($\pm$0.034)	0.705($\pm$0.048)	**0.698**($\pm$0.045)	0.861($\pm$0.014)	0.868($\pm$0.014)
nMIL$^\Omega$	0.687($\pm$0.038)	0.680($\pm$0.045)	0.713($\pm$0.028)	0.687($\pm$0.038)	0.871($\pm$0.013)	0.879($\pm$0.014)

(a) Argentina   (b) Brazil   (c) Mexico

Figure 3: Forecasting evaluation on 3 countries with respect to F1 score for **SVM** , **rMIL^{nor}** , **rMIL^{avg}** and **nMIL** . X-axis is the number of historical days used in the training process. Y-axis shows the average F1 score of 10 runs of experiments.

models performed well consistently across different countries with different number of history days.

6.1.2 How early can nMIL forecast?

In order to study the changes of performance with and without the nested structure, we show the F1 score with varying lead times and historical days from 1 to 5 for **rMIL^{avg}** and **nMIL$^\triangle$** models in Table 3, respectively. We observe that with larger lead time (i.e., forecasting earlier than later), the **nMIL** model does not necessarily lose forecasting accuracy, but is sometimes even better. This can be explained by the fact that several times protests are planned a few days in advance and that civil unrest unfold as a series of actions taken by multiple participating entities over a sequence of days. As the lead time increases, F1 score for forecasting initially drops and then increases back. This behavior is also noted in prior work by Ramakrishnan et. al [19], which includes protest related data from these countries. In comparison to the **nMIL** model, the **rMIL^{avg}** approach, (which collapses the sequential structure encoded within the history of days) seems to perform inconsistently with increasing lead time.

6.2 Do the precursors tell a story?

Quantitative Evaluation. Figures 4a and 4b show the distribution of the estimated probabilities for instances within positive and negative super bags for Argentina and Mexico, respectively.
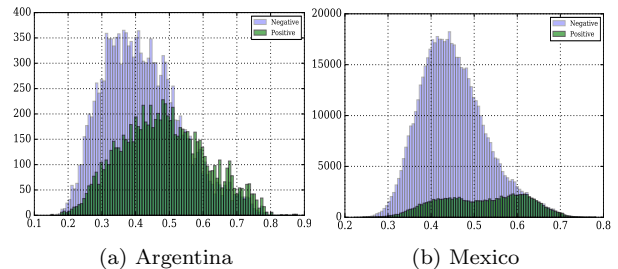

(a) Argentina   (b) Mexico

Figure 4: Estimated probabilities for negative examples (purple) and positive examples (green) for Argentina and Mexico.

The instances within the negative super bags show lower probability estimates by the proposed model and the instances within the positive super bags show higher probability estimates. For Mexico, fewer instances within the positives are assigned high probabilities indicating strength of the proposed model to identify and rank the precursors.

Relative cosine similarity is computed as the pairwise normalized cosine similarity, scaled relative to each event. Figures 5a and 5b show the average cosine similarity value for the precursor documents (probability estimate greater than 0.7) with the target GSR documents.

For Argentina, we observe that on average, the documents on day 5 have the highest semantic similarity to the target event documents (GSR). The documents on day 3 and day 10 have lower similarity compared to the target event.

Table 3: F1-measure for **rMIL^{avg}** and **nMIL$^{\Delta}$** models on Argentina, Brazil, and Mexico with history days from 1 to 5.

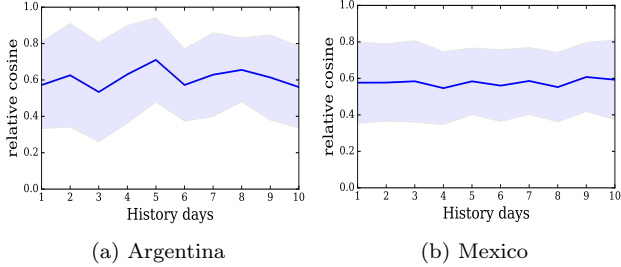| Country		Argentina					Brazil					Mexico				
History Days		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Leadtime 1	**rMIL^{avg}**	0.719	0.714	0.690	0.710	0.705	0.717	0.692	0.696	0.662	0.680	0.815	0.803	0.789	0.796	0.873
	nMIL$^{\Delta}$	**0.745**	**0.735**	**0.722**	0.691	**0.716**	**0.734**	**0.768**	**0.721**	**0.735**	**0.717**	**0.842**	**0.868**	**0.863**	**0.884**	**0.884**
Leadtime 2	**rMIL^{avg}**	0.659	0.624	0.554	0.665	0.500	0.695	0.651	0.573	0.672	0.565	0.846	0.875	0.860	0.878	0.912
	nMIL$^{\Delta}$	**0.664**	**0.675**	**0.740**	**0.710**	**0.751**	**0.699**	0.611	**0.738**	0.639	**0.721**	0.825	**0.889**	**0.914**	**0.909**	0.886
Leadtime 3	**rMIL^{avg}**	0.674	0.606	0.622	0.543	0.578	0.694	0.682	0.620	0.715	0.622	0.819	0.787	0.808	0.750	0.853
	nMIL$^{\Delta}$	0.649	**0.669**	0.560	**0.669**	**0.737**	0.687	0.639	**0.674**	**0.717**	**0.742**	**0.856**	**0.903**	**0.884**	**0.909**	**0.900**
Leadtime 4	**rMIL^{avg}**	0.656	0.558	0.588	0.556	0.476	0.729	0.712	0.720	0.628	0.621	0.809	0.822	0.798	0.878	0.772
	nMIL$^{\Delta}$	**0.676**	**0.693**	**0.670**	**0.712**	**0.631**	**0.754**	0.584	**0.736**	**0.735**	**0.725**	**0.872**	**0.888**	**0.894**	**0.916**	**0.874**
Leadtime 5	**rMIL^{avg}**	0.669	0.676	0.590	0.567	0.575	0.710	0.588	0.616	0.548	0.570	0.828	0.845	0.810	0.733	0.889
	nMIL$^{\Delta}$	0.626	0.676	**0.687**	**0.773**	**0.737**	0.683	**0.665**	**0.657**	**0.697**	**0.735**	**0.833**	**0.937**	**0.878**	**0.935**	**0.931**

(a) Argentina (b) Mexico

Figure 5: Mean of relative cosine values w.r.t target events in history days for Argentina and Mexico.

In order to investigate the relationship between the semantic similarity and the estimated probability by the proposed models, we compare the distribution of relative cosine similarity and relative entity hit score of the precursor documents with the target GSR documents with respect to bag of words features. Entity words in each news document are extracted by an enrichment tool for natural language processing. The relative entity hit score is calculated as the the intersection of entity set of precursor document and the target event divided by the relative minimum length of these two sets.

Figures 6a and 6b show the fitted Gaussian distribution of relative cosine similarities for all documents (green lines) and precursor documents (blue lines) for Argentina and Mexico, respectively. Figures 6c and 6d show the distribution of relative entity hit score for Argentina and Mexico, respectively. These distribution figures demonstrate that the proposed model assigns higher probability to news articles with higher semantic similarity to the GSR articles representing the protests events. These results show the strength of our proposed models in identifying the precursor articles.

Case Studies. We present findings about the identified precursors based on the probability estimate by **nMIL** across three observed protests. In Figure 1, we present a protest event against government in Argentina, and the selected precursors before its occurrence with their estimated probabilities. The titles of news reports as precursors are shown in the timeline.

In Figures 8a and 8b, we present story lines by precursors that were discovered for two different protest events in Argentina and Mexico, respectively. Figure 8a showcases the story line about a protest event in Argentina in December 2014. In this case, the police were protesting against government for better salaries. Before this event, clashes between police and gendarmerie (military policy) had occurred leading to the involvement of several policemen from different parts of the country. The text from news articles

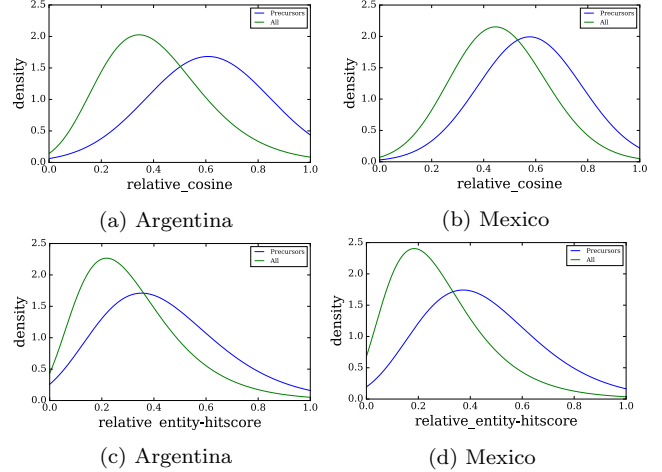(a) Argentina (b) Mexico

(c) Argentina (d) Mexico

Figure 6: The figures on top depict the distribution of relative cosine similarity for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7. The figures in the bottom row depict the distribution of relative entity hit score for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7.

demonstrate the tense situation between the police and government in La Pampa, Argentina identified as precursors.

Figure 8b shows another story line of a continuous protest event in Mexico regarding the infamous case of 43 missing students [2]. The resulting outrage triggered constant protests which were identified by our proposed model. The figure shows a timeline of how the events turned violent leading up to the burning of congressional offices and depicts how different communities joined the movement.

6.3 Can nMIL forecast event populations?

We also evaluated the performance of our **nMIL** approaches for predicting the event populations by solving a multi-class classification problem. In Table 4 we depict the weighted-average F1 score for event populations (here, with categories such as *Government, Wages, Energy, Others* drawn from the GSR). Due to space limitations, we only depict the performance of weighted average F1 score on event population across 1 to 5 historical days with lead time of 1.

The proposed multi-class **nMIL** model outperforms the multi-class **rMIL^{avg}** model. On average, for event population, **nMIL** outperformed **rMIL^{avg}** by 10.5% and 10.6% for Argentina and Mexico, respectively.

[2]https://en.wikipedia.org/wiki/2014_Iguala_mass_kidnapping

Table 4: Multi-Class F1-Measure for **rMIL^{avg}** and **nMIL** models on Argentina and Mexico with historical days from 1 to 5.

	History Days	1	2	3	4	5	Average(Variance)
Argentina	rMIL^{avg}	0.512	0.512	0.473	0.417	0.457	0.474(1e-3)
	nMIL	0.523	0.552	0.515	0.485	0.537	**0.524(7e-4)**
Mexico	rMIL^{avg}	0.576	0.526	0.447	0.547	0.493	0.518(3e-3)
	nMIL	0.570	0.583	0.560	0.615	0.545	**0.575(7e-4)**

6.4 How sensitive is nMIL to parameters?

(a) β in **nMIL** (b) λ in **nMIL**

Figure 7: Sensitivity analysis on β and λ. x-axis represents the varying values for the parameter and y-axis denotes the test accuracy.

There are three main parameters in the proposed **nMIL** model, which are the regularization parameter λ, weight for super bag loss β and threshold for instance level hinge loss m_0. Figures 7a and 7b illustrate the performance of the proposed **nMIL** by varying β and λ, respectively. The test accuracy for different values of λ and β is relatively stable.

7. CONCLUSION AND FUTURE WORK

This paper has presented a novel extension of the multi-instance learning framework for event forecasting and for identifying precursors for protest events. Most existing multi-instance approaches solve problems in object detection in images, drug activity prediction or identify sentimental sentences in text reviews. In contrast, we provide a novel application of MIL algorithms that require a two-level nested structure for event forecasting and precursor modeling.

We have also studied the strengths of our developed methods on open source news datasets from three Latin American countries. Through extensive evaluation and analysis, we illustrate the strong forecasting performance of the proposed methods with varying lead time and historical data. We also show qualitatively via several case studies, the richness of the identified precursors for different protests across different cities. In the future, we plan to incorporate heterogeneous data sources like social media streams for event forecasting within the developed framework. We will also extend our nested multi-instance learning framework by exploring regularized multi-task learning approaches for enforcing similarity of learned parameters, while enforcing spatial and temporal constraints.
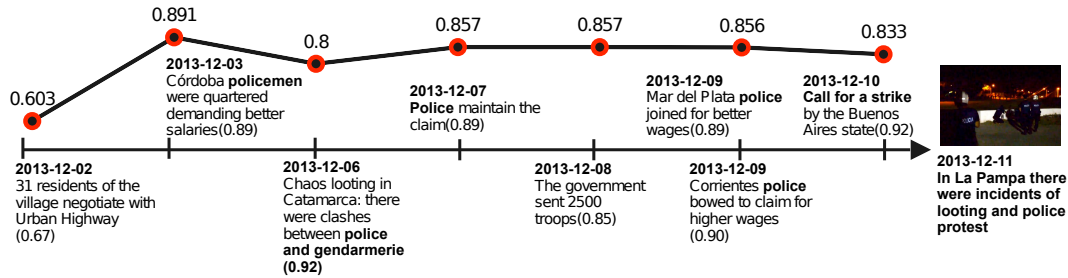
Acknowledgments

8. REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707, April 2011.

[2] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 561–568, 2002.

[4] M. Arias, A. Arratia, and R. Xuriguera. Forecasting with twitter data. *ACM Trans. Intell. Syst. Technol.*, 5(1):8:1–8:24, Jan. 2014.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, Mar. 2003.

[6] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[7] R. Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, July 1997.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, Sept. 1995.

[9] T. Gartner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 179–186, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[10] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1387–1393. AAAI Press, 2013.

[11] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan. Storytelling in entity networks to support intelligence analysts. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1375–1383, New York, NY, USA, 2012. ACM.

[12] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 597–606, New York, NY, USA, 2015. ACM.

[13] S. Laxman, V. Tankasali, and R. W. White. Stream prediction using a generative model based on frequent episodes in event sequences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 453–461, New York, NY, USA, 2008. ACM.

[14] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[15] G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. In S. C. H. Hoi and W. L. Buntine, editors, *ACML*, volume 25 of *JMLR Proceedings*, pages 253–268. JMLR.org, 2012.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[18] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment

(a) A continuous police protest in Argentina against government for better salary. In the beginning, policemen at Cordoba were requesting for better salaries. Later on, police in Catamarca were involved in clashes with gendarmerie. Three days before the target event, the government sent out troops and more and more police joined for the same purpose. One day before the event, Buenos Aires state call for a strike.

(b) Protester in Mexico burned the congressional offices for justice for the missing students. In the beginning, students were marching for justice. Gradually, more communities such as artistic and policing community joined the event. Later on, children, youth, adults, students and teachers blocked traffic for protest.

Figure 8: Case Studies

to public opinion time series. In W. W. Cohen and S. Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[19] N. Ramakrishnan, P. Butler, S. Muthiah, and et al. "beating the news" with embers: Forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1799–1808, New York, NY, USA, 2014. ACM.

[20] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *In Proceedings of the 1st International Workshop on Mining Social*, 2009.

[21] Y. Rong, H. Cheng, and Z. Mo. Why it happened: Identifying and modeling the reasons of the happening of social events. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1015–1024, New York, NY, USA, 2015. ACM.

[22] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[23] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from twitter posts. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'12, pages 231–238, Berlin, Heidelberg, 2012.

Springer-Verlag.

[24] X. Wang, Z. Zhu, C. Yao, and X. Bai. Relaxed multiple-instance SVM with application to object discovery. *CoRR*, abs/1510.01027, 2015.

[25] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 468–479, 2003.

[26] L. Zhao, F. Chen, C. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 963–971, 2015.

[27] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1503–1512, New York, NY, USA, 2015. ACM.

[28] Z.-H. Zhou and J.-M. Xu. On the relation between multi-instance learning and semi-supervised learning. In Z. Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 1167–1174. ACM, 2007.